

Extending modularity by incorporating distance functions in the null model

Xin Liu,^{1,2,3} Tsuyoshi Murata,⁴ and Ken Wakita^{1,*}

¹*Department of Mathematical and Computing Sciences, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro, Tokyo, 152-8552 Japan*

²*CREST, Japan Science and Technology Agency
K's Gobancho, 7, Gobancho, Chiyoda, Tokyo, 102-0076 Japan*

³*Department of Mathematics, Wuhan University of Technology
122 Luoshi Road, Wuhan, 430070 China*

⁴*Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro, Tokyo, 152-8552 Japan*

(Dated: October 17, 2012)

Modularity is a widely used measure for evaluating community structure in networks. The definition of modularity involves a comparison of the fraction of within-community edges in the observed network and a null model. In the original definition the null model only considers the node degree to rewire edges randomly, failing to be a good representation of many real-world networks. To handle this problem, we incorporate distance functions in the null model to facilitate edges between certain nodes while respecting the degree factor. This enables us to create a framework for generating appropriate modularities adapted to various networks.

A network community is a group of nodes, within which edges are dense, but between which edges are sparse [1]. In practice, optimization methods are widely used for detecting communities in networks [2]. The basic idea is to define a quantity measure for evaluating the "goodness" of a partition of the observed network into communities, and then to search through possible partitions for the one with the highest score. A variety of partition measures have been proposed, but the most famous one is known as the modularity [3]. Formally, modularity is defined to be the fraction of edges within communities in the observed network minus the expected value of that fraction in a null model, which serves as a reference and should characterize some features of the observed network. In a mathematical expression, modularity reads

$$Q(\mathcal{C}) = \frac{1}{2m} \sum_{i,j=1}^n (A_{ij} - P_{ij}) \delta(l_i, l_j), \quad (1)$$

where \mathcal{C} is a partition represented as a community-assignment vector on the right-hand side of the equation, with element l_i indicating the community membership of the i th node v_i , n is the number of nodes, m is the number of edges, A_{ij} is the element of the adjacency matrix \mathbf{A} representing the number of edges between v_i and v_j in the observed network, P_{ij} is the expected value of that number in the null model, and $\delta(\cdot, \cdot)$ is the Kronecker's delta.

In the original version of modularity proposed by Newman and Girvan (NG modularity) [3], the null model preserves the degree sequence of the observed network and rewires edges randomly, giving [4, 5]

$$P_{ij}^{\text{NG}} = k_i k_j / 2m, \quad (2)$$

where $k_i = \sum_{j=1}^n A_{ij}$ is the degree of v_i . Note that in this null model the number of edges between nodes only de-

pends on their degrees. However, this is not always the case for the network under observation, because other factors such as the distance between nodes can strongly affect their connections. For example, in spatial networks [6] such as the Internet, road networks, and flight connections, long distance connections are always restricted due to financial cost or physical constraints. In social networks, individuals with a shorter distance with respect to their interests are more likely to be connected. Consequently, this null model may fail to be a valuable reference and result in a less accurate partition measure.

We develop a new null model^{Note*}, which takes the distance factor into consideration. In our null model, the number of edges between v_i and v_j is rewired according to the probability

$$P_{ij}^{\text{dist}} = (\tilde{P}_{ij} + \tilde{P}_{ji}) / 2, \quad (3)$$

$$\tilde{P}_{ij} = \frac{k_i k_j e^{-(d_{ij}/\sigma)^2}}{\sum_{t=1}^n k_t e^{-(d_{ti}/\sigma)^2}}, \quad (4)$$

where $\sigma \in (0, +\infty)$ is a field range parameter which will be explained later, $d_{ij} \geq 0$ is the distance between v_i and v_j .

Eq.(4) can be interpreted by the data field idea proposed by Li [9], which introduces the field theories in Physics for describing interactions between particles into the data space. Now suppose each node exerts forces on others by generating a field. The field theories (including the gravitational field, the electrostatic field, the magnetostatic field, and etc.) say that the potential at a point

^{Note*}In this paper we focus on undirected and unweighted networks. Developments for directed and/or weighted networks can be easily extended as in [7, 8].

in space is directly proportional to the power of the field source (such as the mass or charge), and decreases as the distance to the source increases. Based on this, we can calculate the potential at v_j of the field driven by v_i as

$$\varphi_i(j) = k_i e^{-(d_{ij}/\sigma)^2}. \quad (5)$$

where k_i is supposed to be the power of v_i , and the distance function $f(d) = e^{-(d/\sigma)^2}$, falling in $(0, 1]$, monotonically decreases with d . The parameter σ reflects the interaction range of the field — If σ is small, $f(d)$ decreases sharply, indicating a short-range field; If σ is large, $f(d)$ decreases slowly, indicating a long-range field. Note that the fields driven by different nodes can be superposed. Suppose the potential is a scalar quantity without direction. The superposed potential at v_j is

$$\varphi_{\text{sup}}(j) = \sum_{t=1}^n \varphi_t(j) = \sum_{t=1}^n k_t e^{-(d_{tj}/\sigma)^2}. \quad (6)$$

Then Eq.(4) can be rewritten as

$$\tilde{P}_{ij} = \frac{\varphi_i(i) \varphi_j(i)}{\varphi_{\text{sup}}(i)}, \quad (7)$$

where we have used $d_{ii} = 0$ and $d_{ij} = d_{ji}$. That is, \tilde{P}_{ij} is the the product of potentials at v_i of the fields driven by v_i and v_j separately, divided by the superposed potential at v_i .

In the following, we describe the features of our null model. First, it can be found that

$$P_{ij}^{\text{dist}} = P_{ji}^{\text{dist}}, \quad (8)$$

which implies that edges in our null model are undirected. Second, it is easy to derive that

$$\sum_{i,j=1}^n \tilde{P}_{ij} = \sum_{i,j=1}^n \tilde{P}_{ji} = \sum_{i=1}^n k_i = 2m, \quad (9)$$

and hence

$$\sum_{i,j=1}^n P_{ij}^{\text{dist}} = 2m. \quad (10)$$

That is, the number of edges of the observed network is preserved. Third, Eq.(4) tells us that \tilde{P}_{ij} is positively related to k_i and negatively related to d_{ij} . In other words, connections tend to link to high degree nodes and nodes with short distances.

Based on our null model, we can define distance modularity (dist-modularity) as

$$Q^{\text{dist}}(\mathcal{C}, \sigma) = \frac{1}{2m} \sum_{i,j=1}^n (A_{ij} - P_{ij}^{\text{dist}}) \delta(l_i, l_j). \quad (11)$$

From Eq.(10) and (11), it is clear that $Q^{\text{dist}} \in [-1, 1]$, the same as NG modularity.

Note that there is a range parameter $\sigma \in (0, +\infty)$ in P_{ij}^{dist} , hence different σ brings different dist-modularity. In one extreme case, i.e. $\sigma \rightarrow 0+$, the range of the field driven by each node is so short that the potential only exists at the source node. This gives

$$\lim_{\sigma \rightarrow 0+} P_{ij}^{\text{dist}} = \begin{cases} k_i, & \text{if } i = j; \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

and hence

$$\lim_{\sigma \rightarrow 0+} (A_{ij} - P_{ij}^{\text{dist}}) = A_{ij} - k_i \delta(i, j), \quad (13)$$

As a result, the Laplacian matrix \mathbf{L} , the key matrix in graph partitioning [10] and spectral clustering [11], and the modularity matrix \mathbf{B} , defined to be $\mathbf{A} - \mathbf{P}$ [4, 5], are unified as σ approaches 0 — The only difference is the sign.

In the other extreme case, i.e. $\sigma \rightarrow +\infty$, the range of the field driven by each node is so long that the potential at each node equals. This gives

$$\lim_{\sigma \rightarrow +\infty} P_{ij}^{\text{dist}} = k_i k_j / 2m. \quad (14)$$

As a result, dist-modularity reduces to NG modularity.

It is interesting to note that as σ increases from 0 to $+\infty$, optimizing dist-modularity brings community structure at different scales, from coarse to fine. First, $\lim_{\sigma \rightarrow 0+} Q^{\text{dist}}$ can be rewritten as

$$\lim_{\sigma \rightarrow 0+} Q^{\text{dist}}(\mathcal{C}, \sigma) = \frac{1}{2m} \sum_{g=1}^c \sum_{i,j \in C_g} A_{ij} - 1, \quad (15)$$

where c is the number of communities, C_g is the g th community. Eq.(15) implies that $\lim_{\sigma \rightarrow 0+} Q^{\text{dist}}$ is optimized to 0 when the network is divided into only one community or several communities corresponding to its connected components. Obviously this is the community structure at the coarse scale. Second, optimizing $\lim_{\sigma \rightarrow +\infty} Q^{\text{dist}}$, i.e. optimizing NG modularity results in community structure at the fine scale. Third, as σ ranges from 0 to $+\infty$, optimizing Q^{dist} brings multi-scale community structures that fall between the above two extremes.

One issue that has not been addressed is the distance between nodes. For plain networks with only link information on nodes, d_{ij} can be calculated from the i -th and j -th rows of the adjacency matrix \mathbf{A} . Many networks in real world have additional information on nodes (node attributes), such as the geographical position of a location or the profile of a person. For these networks, d_{ij} can be calculated from the attributes of v_i and v_j using a specific vector distance measure, such as the Euclidean distance, the Manhattan distance and the Minkowski distance.

So far we have proposed a new null model by incorporating a distance function. In the following, we show

how this null model can be generalized to produce a family of dist-modularities for various networks. Now let us look back at Eq.(5) and (6), the expression of the potentials, which is the core component of our null model. It can be found that the power of the field source and the distance function are specifically specified as the node degree and $e^{-(d/\sigma)^2}$, respectively. However, we have many more choices while preserving the desired features of the null model. Suppose the power of v_i and the distance function are denoted by N_i and $f(d)$, respectively. Then

$$\varphi_i(j) = N_i f(d_{ij}), \quad (16)$$

$$\varphi_{\text{sup}}(j) = \sum_{t=1}^n N_t f(d_{tj}). \quad (17)$$

Substituting Eq.(16) and (17) into Eq.(7), we have

$$\tilde{P}_{ij} = \frac{N_i N_j f(d_{ij})}{\sum_{t=1}^n N_t f(d_{ti})}. \quad (18)$$

Finally, with Eq.(3) and (18) we can derive that

$$\sum_{i,j=1}^n P_{ij}^{\text{dist}} = \sum_{i=1}^n N_i. \quad (19)$$

Eq.(19) implies that the number of edges of the observed network is preserved if $\sum_{i=1}^n N_i = 2m$ is satisfied. This can be easily achieved by normalizing N_i as

$$N_i = \frac{N_i}{\sum_{i=1}^n N_i} 2m. \quad (20)$$

Due to the above developments, there is a large freedom in specifying N_i and $f(d)$. For example, N_i can be equally distributed among all nodes, i.e. specifying $N_i = 2m/n$. In networks with node attributes, N_i can be specified as the most representative attribute, or a comprehensive index obtained from several attributes. As for $f(d)$, the common choices are

$$f(d) = e^{-(d/\sigma)^r}, \quad (21)$$

$$f(d) = 1/(1 + (d/\sigma)^r), \quad (22)$$

$$f(d) = 1, \quad (23)$$

$$f(d) = \begin{cases} 1, & \text{if } d \leq \sigma; \\ 0, & \text{otherwise,} \end{cases} \quad (24)$$

where $\sigma \in (0, +\infty)$ and $r > 0$ are parameters. Note that functions (21)-(24) are monotonically decreasing in the domain $d \in [0, +\infty)$. However, this is not a necessary constraint. For example, $f(d)$ can be learned from the observed network as [12]

$$f(d) = \left(\sum_{\substack{i,j=1 \\ d_{ij}=d}}^n A_{ij} \right) / 2m, \quad (25)$$

with a binning procedure to smooth the function. $f(d)$ can even be replaced by a similarity function that monotonically increases with d .

The freedom in specifying N_i and $f(d)$ enables us to create a framework to produce a family of dist-modularities. Within this framework, it is interesting to see some relations with previous work. For example, 1) When $N_i = 2m/n$ and $f(d) = 1$, we have $P_{ij}^{\text{dist}} = 2m/n^2$, indicating the Erdős-Rényi random graph [13]; 2) When $N_i = k_i$ and $f(d) = 1$, we have $P_{ij}^{\text{dist}} = k_i k_j / 2m$, indicating the null model of NG modularity.

In conclusion, we incorporate distance functions in the null model to capture the features of real-world networks. Taking this null model as a reference for comparing the fraction of within-community edges with the observed network, we create a framework for generating a family of dist-modularities adapted to various networks, including networks with node attributes. In addition, we have several interesting findings within this framework as below.

- Laplacian matrix and modularity matrix can be unified, providing an in-depth view of the close relationship between graph partitioning/spectral clustering and community detection;
- NG modularity can be exactly recovered as a special case of dist-modularity.
- Dist-modularity can be used to detect communities at different scales, from coarse to fine.

RELATED WORK

The study of community detection in networks has a long history. It is closely related to graph partitioning [10] in computer science, and hierarchical clustering [14] in sociology. In the past decade, this study has attracted a great deal of interests and various methods were proposed [2, 15–18]. In particular, modularity optimization [5, 19–26] is widely used despite its intrinsic limits [27–29].

Recently, some researchers considered community detection in networks with node attributes. Expert et al. [12] and Cerina et al. [30] proposed methods by factoring out the effect of space in spatial networks where the geographical information on nodes is available [6]. Yang et al. devised a discriminative approach for combining the edge and node attributes [31].

ACKNOWLEDGMENTS

This work was partly funded by CREST, JST and by NSFC under grant number 61203154.

* To whom correspondence should be addressed.

E-mail: wakita@is.titech.ac.jp

- [1] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
- [2] S. Fortunato, Physics Reports **486**, 75 (2010).
- [3] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
- [4] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).
- [5] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
- [6] M. Barthélemy, Physics Reports **499**, 1 (2011).
- [7] E. A. Leicht and M. E. J. Newman, Phys. Rev. Lett. **100**, 118703 (2008).
- [8] M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004).
- [9] D. Li and Y. Du, *Artificial Intelligence with Uncertainty* (Chapman and Hall/CRC, London, 2007).
- [10] B. Kernighan and S. Lin, Bell Syst. Tech. J. **49**, 291 (1970).
- [11] U. V. Luxburg, Statistics and Computing **17**, 395 (2007).
- [12] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, Proc. Natl. Acad. Sci. USA **108**, 7663 (2011).
- [13] P. Erdős and A. Rényi, Publ. Math. **6**, 290 (1959).
- [14] J. Scott, *Social Network Analysis: A Handbook*, 2nd ed. (Sage Publications, Newberry Park, CA, 2000).
- [15] M. E. J. Newman, Nature Physics **8**, 25 (2011).
- [16] L. Danon, J. Duch, A. D.-Guilera, and A. Arenas, J. Stat. Mech., P09008 (2005).
- [17] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 056117 (2009).
- [18] J. Leskovec, K. J. Lang, and M. W. Mahoney, in *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, NC, USA, 2010) pp. 631–640.
- [19] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
- [20] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
- [21] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
- [22] A. Medus, G. Acuna, and C. O. Dorso, Physica A **358**, 593 (2005).
- [23] P. Schuetz and A. Cafilisch, Phys. Rev. E **77**, 046112 (2008).
- [24] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech., P10008 (2008).
- [25] K. Wakita and T. Tsurumi, in *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada, 2007) pp. 1275–1276.
- [26] X. Liu and T. Murata, Physica A **389**, 1493 (2010).
- [27] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007).
- [28] A. Lancichinetti and S. Fortunato, Phys. Rev. E **84**, 066122 (2011).
- [29] B. H. Good, Y.-A. de Montjoye, and A. Clauset, Phys. Rev. E **81**, 046106 (2010).
- [30] F. Cerina, V. D. Leo, M. Barthelemy, and A. Chessa, PloS one **7**, e37507 (2012).
- [31] T. Yang, R. Jin, Y. Chi, and S. Zhu, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Paris, France, 2009) pp. 927–936.